THE LATEST REVISION OF THE BINET INTELLIGENCE TESTS

By CYRIL BURT, M.A., D.Sc.

NEARLY twenty-five years ago, when the Binet-Simon tests were first coming into vogue, I endeavoured to examine, in two articles in this REVIEW, their merits and limitations as a means of measuring innate intelligence.* Since then they have been used throughout the country by teachers, psychologists, and school medical officers to test the abilities of pupils and to discover the dull, the backward, the potential scholarship winners, and, most important of all, those who are certifiably defective. Meanwhile, psychologists themselves have made several attempts to improve the scale, and to eliminate, so far as possible, the obvious defects of the earlier version.

The Old Scales

In London, with Binet's express permission and in consultation with his colleague, Dr. Simon, an investigation was started by the Education Department of the County Council with a view to restandardizing the original French arrangement for use with English children. This was finally published in a *Handbook of Tests for Use in Schools*,† and made the basis for a survey of the distribution of intelligence in typical areas of the County. During the same period, Professor Terman in California was making an independent adaptation of the scale for American schools. Although both English and American revisers had been working with little or no knowledge of the detailed findings of each other, it turned out that both had been led to make very much the same changes. There were, however, a few big differences. In particular, Terman added a number of ingenious and original tests at the upper end of the scale for the examination of brighter children and adults. In London, therefore, a second investigation was started to examine, and if necessary restandardize, this further version. Professor Terman generously gave permission to publish an English adaptation; but added that he was contemplating a complete re-investigation of the tests himself. Accordingly, the full adaptation of the first Terman version was only issued privately to clinics; the most essential changes—the new assignment of ages—were published as an appendix to the *Joint Report* of the Board of Education and the Board of Control on Mental Deficiency.

The New Scale

Terman's new revision is now at length available;* and those whose work entails the assessment of intelligence are inquiring whether the new version is, as it claims to be, genuinely superior to the older versions, and whether it has successfully eliminated the defects that had become obvious in the first scale without destroying any of the merits. A large committee, comprising all the leading educational psychologists in Great Britain, together with representatives or members of every body or institution engaged in mental tests, has been set up to work through the wording of the new revision, test by test, and to check the age standardizations on the basis of extensive experiments all over the country. A provisional "translation" (if I may so call it) is now available, and may be obtained from the Psychological Department,

---

* "The Measurement of Intelligence by the Binet Tests," EUGENICS REVIEW, 1914, vi, 36 and 140.

† P. S. King & Son, 1923. The results of our experiments were summarized from time to time in the *Reports of the Psychologist to the Council*. The final standardization was issued by the L.C.C. in a long official *Report* entitled "Mental and Scholastic Tests" in 1921; and the new age-assignments were also appended to the Board of Education's *Report of Tests of Educable Capacity* (1924, pp. 200-3).

* L. M. Terman and Maud A. Merrill, *Measuring Intelligence*, G. G. Harrap & Co., 1937.

University College, London, for private experimental use.

Thanks to the prompt co-operation of teachers, psychologists and medical officers, a large mass of data has already come to hand; and it may be of interest to summarize briefly the chief conclusions that emerge from a first preliminary survey of the results. To the eugenist perhaps the most interesting outcome will be the light thrown by such data upon the distribution of innate intelligence among the general population and the changes in the mental level (if any) revealed by the present survey as compared with those carried out twenty and twenty-five years ago. But before any answer to such questions can be sought, we must scrutinize more closely the validity of the scale as such. The old set of tests formed a convenient and practical method for quick clinical diagnosis rather than a reliable scientific instrument for statistical surveys. And the publication of the new version has revived the ancient criticisms in an even more emphatic form;* "the prolonged worship of the Binet scale," says Dr. Cattell, "has left us with an encumbering heritage of erroneous conceptions, especially in matters concerning the distribution of intelligence and its role in society." So far the dispute has turned largely on *a priori* principles, but is none the less valuable for that; here I propose to rely mainly on actual figures.

The Main Differences

As regards the chief practical purpose of the scale—the diagnosis of dull and defective children—there seems little question that the new revision is decidedly more efficient than the old. As regards the wording and the age-assignments of the numerous tests that have been retained, the new revision frequently accepts the modifications proposed in the former London version. The principle of "internal grading," too, which was advocated in this REVIEW, has been far more freely used—notably, in the vocabulary test. Many of the problems, however, are entirely new; and here the American age-assignments and even the American wording seem often far from appropriate to English children. The second "Paper Cutting test," which Terman assigns to the third or highest level of "Superior Adults," we find can be done at age 14; on the other hand, "Giving Similarities between three Things," which cannot be done by the average Londoner until age 14, Terman assigns to age 11. Between the ages of 4 and 14, out of 66 tests 32 would appear to be misplaced by at least a year. It would, therefore, be unwise for teachers, medical officers, or field workers to commit themselves rigidly or finally to the scale until the requisite modifications are known.

The Need for Item-Analysis

The ideal plan would be to take each separate test-problem, and examine its special value as a criterion of intelligence. Curiously enough, this has rarely been attempted: no doubt, the labour of separately evaluating eighty or ninety (or with the new scale 130) tests deters the eager investigator. With the original Binet scale we worked out, both for defectives and for normal children, a "coefficient of colligation" between each separate test and intelligence as assessed by competent teachers.* This was, I believe, the first experiment in what is now called "item-analysis"; but, since those early efforts, the mathematical technique has been considerably refined. In the revised scale, most of the poorer tests that figured in the original scale (e.g., "Suggestion," "Months," "Age," "Sex," "Surname") have silently been dropped. The newer tests inserted in their place are based on accepted psychological principles; nevertheless those principles have really been derived more from group testing than from individual examinations. An intensive item-analysis for each test separately is thus an urgent requisite. Such an investigation, however, will call for a long and patient research. Meanwhile, we may usefully attempt a more general

* P. E. Vernon, "The Stanford Binet Test as a Psychometric Method"; R. B. Cattell, "Measurement versus Intuition in Applied Psychology," *Character and Personality*, 1937, vi, 99.

* L.C.C. 1921 *Report*, Table XXXI.

estimation of the merits of the new revision as a whole.

As I pointed out in my original examination of the scale, its difficulties and defects arise largely from the adoption of the plan of "external" instead of "internal grading." This plan proves convenient for the practical examiner; but makes it far from easy to test the tests themselves. With "internally graded" tests we can arrange the persons tested in order according to their ability in each test, and so correlate the tests with each other or with an independent criterion in the ordinary way. With "externally graded" tests, we can only say whether each person passes or fails, and so at best estimate the correlation between test and intelligence from fourfold tables by a "coefficient of colligation" or the like—never a very satisfactory procedure.

Variability of the Order of Difficulty

With an externally graded scale, like the Binet-Simon series, everything turns upon the relative difficulty of the test-problems. The standardization of each problem in terms of a mental age assumes that the order of difficulty is constant for the two sexes, for different social classes, for different ages, for different types of child, and above all for different localities. Thus, if a child repeats four numbers backwards, Terman would give him a mental age of 9; if he repeats six numbers forwards, i.e., in the order in which they have been recited to him, he would get a mental age of 10. Now with London children it is found that the latter test is actually easier than the former. And so with many other tests: the order of difficulty is often reversed. Worse still, when we experiment with the scale as a whole, there seems to be no fixed order at all: what is easier for one child may be harder for another. Indeed, the early critics of the scale were constantly pointing out how no two editors of Binet's scale ever agreed over the relative difficulty of the several tests. The orders of difficulty seem to vary with different examiners as well as with different examinees.

At the very outset, therefore, in examining the validity of the whole proposal, the task of the psychologist must be to compare these different orders, and see whether they vary so widely as to invalidate the very foundations of the scale. The lay critic is generally content to exhibit one or two glaring examples of discrepancy, and base his argument on those. The scientific investigator endeavours to assess the amount of agreement or disagreement shown over the scale in its entirety. He sums the discrepancies between the orders to be compared, and so obtains, by a simple formula and in a single figure, an exact measure of the total disagreement. This means, to use a familiar distinction, that his statistical evaluation must depend mainly on "correlating persons" instead of "correlating tests." The results are expressed in terms of what may be called "consistency coefficients." *

* The method is by no means new, though it has recently given rise to much discussion. There are several minor difficulties, which, however, can very easily be met. When (as here) we start with a rectilinear "order of difficulty" instead of with a normally distributed set of marks, we have to base our coefficient of correlation on the displacements in the orders, i.e. the "rank-differences": must we square them (according to the product moment formula) or should we leave them unsquared (according to Spearman's much-criticized "footrule")? Now, when we correlate persons the first or "total" correlations, calculated by the product moment formula, often appear exceedingly high—so high as to be somewhat misleading to those who are more familiar with correlations between tests; with the rank-squaring method, extreme divergences are heavily weighted; ·89 (for example) of the "footrule" becomes ·99, and minor variations near this point are apt to be obscured. Hence, in some of my earliest reports, though I used the footrule method introduced by Spearman, I did *not* apply what would generally be regarded as the necessary correction. But, after all, our coefficients are no more than a convenient device—a mode of averaging the differences (and often blurring them), whereas the real point of interest is the differences themselves. Consequently, if we are studying correlations between children, it is both safer and more instructive to rely on the detailed table of rank differences than on a small table of coefficients, however obtained. For the correlations between examiners the rank-squaring method was used.

As some readers might feel a little dubious about the validity of correlating orders for persons, I attempted at the same time (by means of Yule's colligation formula) to estimate the correlations between the tests (*loc. cit.*, Table XXIX). The results are in entire agreement. Moreover, the group-factors found by either method with the Binet tests are similar to those revealed by analysing correlations between *scholastic* tests: e.g. the verbal and non-verbal types stand out discernibly in both (cf. L.C.C. 1917 *Report*, pp. 58-9).

Consistency Coefficients for Examinees

In the earliest inquiries the persons correlated were the examinees. Binet's method, as we have seen, merely requires the examiner to note whether the child "passes" or "fails." Nevertheless, in preliminary trials, it is not difficult, with a full record of the child's performance, to sort the particular tests that have been given to him in order of their apparent difficulty for him, and then compare these orders as obtained from different individuals. As noted in my previous paper, the results with the earliest form of the scale seem decidedly unfavourable, for the correlations were far from perfect. Binet had obviously underestimated the extent to which children vary in *kind* of ability as distinct from *amount*. Of two children possessing the same general intelligence one may be far better at the memory tests, another at the manual tests; one may be quick at perception but too impatient to reflect; another may do well at tests of reflective reasoning, but prove as unobservant as he was unpractical. A formal statistical analysis of the tabulated results disclosed, not only a main "general factor" of intelligence, but also secondary "group factors" classifying the children into groups or "types."

As the number of children ran to hundreds and then to thousands, it became impossible to correlate all the individual orders. They were accordingly correlated by batches: for example, by using *average* orders based on the percentages of failures, we can compare normals with defectives, boys with girls, children from the poorest social classes with children from the best, and children of one nationality with children of others. The group-differences are then discovered to be much smaller than the individual differences.

Sex and Social Differences

This will become clear if we turn for a moment to the data obtained from the original scale. For example, if we first correlate the individual orders obtained from a large number of boys with those obtained from a large number of girls, it appears that the orders of the girls agree much more closely with each other (averaging, when obtained by the footrule method, ·78) than they agree with the orders of the boys (·66): and similarly for the boys. If now we correlate the *average* orders for either sex, the correlation rises to ·867. The correlation between children of superior social status and those coming from the slums was found to be higher still, namely, ·890. On the other hand, the correlation between normal children and defective was only ·849.

At first sight these coefficients may seem high; and so they would be, if they were correlations between tests instead of between persons. Let us therefore examine the orders themselves; it will then be seen that the detailed amount of divergence exhibited by the correlated orders is still quite serious. With only 65 tests, the total of the "rank-differences" amounts to 176 (between the two sexes), 153 (between the two social classes), and 211 (between normals and defectives. On tabulating the discrepancies according to their size and then examining the nature of the tests concerned, we can quickly ascertain what are the chief causes: once again, the tests are evidently revealing, not merely differences in degree of intelligence, but also differences in mental quality or type. The girls, for example, being rather of a verbal type and good memorizers as well, do better at all tests involving memory for words—e.g., reading, dictation, naming colours; the boys do better at the non-verbal or mechanical tests — arranging weights, drawing from memory, counting up coins.*

Now the original Binet scale had a heavy verbal bias; and such a bias is particularly hard upon the mentally defective child, who as every teacher knows, is better at practical things than he is at reading, writing, spelling and rational conversation. This defect in Binet's own series has, of course, not escaped Professor Terman: and, indeed, the chief difference between the former Terman series and the new is that, whereas

* See L.C.C. 1919 *Report*, Tables II and III, and 1921 *Report*, Tables VIII, XXVIII, XXIX; also this REVIEW, *loc. cit.*, pp. 160 *et seq.*

the former was still overweighted with verbal tests, the new is, if anything, overweighted with practical or manual tests.

Consistency Coefficients for Examiners

But the order and age-assignments of the several tests vary not only with the personality of each child, but also with the "personal equation" of the examiner. If, for example, in testing my pupils, I utter my instructions less distinctly than Binet, will not this make the verbal tests seem harder in my series than in Binet's? And however carefully the examiner's procedure is laid down in the manual, there are bound to be differences in technique and approach. How are we to gauge their influence? Once more we "correlate persons": and this time the persons are the examiners. Broadly put, our question is this: are the examiners' orders, though never precisely the same, at least consistent enough for us to assume that one main general factor is operative in all, or do differences of nationality, language, individual procedure and the like deprive the tests of all comparability? Let me once again call to mind the figures obtained with the earlier form of the tests. For the original scale the answer was given by printing a table of intercorrelations. The different examiners had obtained their several results in France, Germany, Italy, America and various parts of Great Britain: yet their average intercorrelation was ·873—i.e., there seemed a high degree of general agreement. The English examiners agreed with each other to the extent of between ·97 and ·99: with the different American examiners to the extent of ·72 to ·94: with Terman himself to the extent of ·93.

The General Factor for Examiners

But this is not enough for the psychologist. He attempts what is called a "factor-analysis" to discover whether or not the agreement is attributable to a single factor running through all: after all, the agreement between the English and American might be due to their common language and the agreement between the French and English to something quite different—e.g., their geographical proximity. This point must also be tested. After making the necessary calculations it was concluded that the table of figures thus analyzed "strongly suggests a single central factor, underlying and determining (though in slightly different proportions) all the various orders." * To put it in a technical fashion, 88 per cent. of the "variance" (i.e., of the variations in the individual figures) is attributable to something common to all examiners, 12 per cent. to specific influences more or less peculiar to a few or possibly only to each particular one. With the new Terman revision agreement is higher still: the correlation between his order and ours (only a provisional one as yet) rises to ·96.

The Influence of Verbal and Non-Verbal Types

On the other hand, it seems as if the contrast between the verbal type of child and the non-verbal will be shown up almost as clearly as before. Teachers and psychologists have frequently argued that, if the Binet scale contains such a strong verbal bias, then it cannot be trusted as a measure of general intelligence. More than one critic has objected that the tests might often send a normal child to the special school simply because he was of a practical rather than of a scholastic type. So, too, on seeing the new Terman revision, many are tempted to ask

* L.C.C. 1921 *Report*, pp. 136-7. The method I originally used was an application of Spearman's "two-factor theory" (devised by Spearman himself for correlations between tests) to correlations between persons. Soon after, Professor Godfrey Thomson strongly criticized Professor Spearman's argument on the ground that it involved the familiar logical fallacy of *illicit conversion*. Let us, therefore, apply Professor Godfrey Thomson's own recent criterion: it turns upon the fact that in a table of a given size it is an arithmetical impossibility for correlations to rise above a certain calculable boundary unless there is a general factor running through all the items correlated. If N is the number of items correlated (here persons), and if n is the number of persons in whom the factor appears, then the average correlation cannot be greater than $\frac{n-2}{N-1}$. Here there were 11 persons correlated; hence $N-1$ is 10. The average correlation is over ·87, i.e. nearly $\frac{9}{10}$. Thus $n-2$ must be at least 8, possibly 9. Consequently we infer that the central factor must go through at least $(8+2)=10$ out of the 11 arrangements, and quite possibly through all of them.

whether it is not a mistake to retain tests of reading and of arithmetic in a scale that is meant to test not learning but innate capacity; and others have inquired whether the new mechanical and manual bias will have the opposite error to the old, and make the scale more suited for discovering children for trade schools than scholarship winners for secondary schools.

The General Factor for Examinees

Once again it is largely a problem in what the psychologist calls "factor analysis." Put into technical terminology the question is this: how much of the total "variance" (i.e., of the variations in performance due to any and every type of cause) is attributable to a single general factor common to all children, and how much is due to specific and irrelevant factors—sex, social opportunity, school teaching, and above all perhaps qualitative difference in mental type?

The data that have already been received enable us to give some first provisional answer to this question. For this preliminary survey the correlations have been calculated, and the coefficients factorized, by the same methods as were used in the original research—namely, on simple "two factor" principles. The following table shows what proportions are attributable (i) to the "common factor" and (ii) to all other influences in the three main London revisions to which I have referred.

The figures for the first two revisions are taken from earlier papers and are based on fairly wide investigations. Those for the last are so far based on but comparatively few cases: but, so far as they go, they indicate

TABLE SHOWING RELATIVE CONTRIBUTIONS OF DIFFERENT FACTORS TO THE VARIATIONS IN THE RESULTS OF THE BINET SCALES

(*Factor-Analysis based on Pooled Correlations between Persons: children aged 8 to 13 only*)

	(i) First or General Factor	(ii) Remaining or Specific Factors
London Revision of Original Binet Scale, 1921	72 per cent.	28 per cent.
Original Terman Scale, 1928	67 per cent.	33 per cent.
New Terman Scale, 1938	76 per cent.	24 per cent.

that the new scale is appreciably more reliable than the old. But far more numerous data are needed.

We should therefore be extremely grateful to any teachers, psychologists, medical officers, and others using the new scale, who would be good enough to co-operate in this preliminary statistical work. After all, the whole diagnostic value of the scale turns on the way we shall ultimately answer this preliminary question: how far is our order of difficulty trustworthy and how far is that order—and consequently our whole scheme of age-assignments—liable to be disturbed by wholly irrelevant conditions?

With a little ingenuity we may perhaps turn the defect into a merit, and even ultimately make the scale do two things at once: measure the child's general intelligence with a rough but reasonable degree of accuracy and at the same time throw side-lights on the kind as well as on the nature and extent of the special abilities or the special defects that he displays.